COMPARISON OF TWO BIVARIATE STRATIFICATION SCHEMES

Alan Roshwalb and Scott Regnier
Georgetown University, Washington, D.C. 20057

1. INTRODUCTION

In this paper, we examine stratified sampling for inventories when at least two variables are available for stratification. The goal of the sampling plan is to estimate the total value of the inventory, Y, or the population mean, $\bar{Y}$. A population's elements are to be apportioned into H strata.

Let's assume that the variables $Z_1>0$ and $Z_2>0$ are available for stratification. It is well known that when only one variable is available for stratification, the Dalenius-Hodges cum $\sqrt{f}$ rule (D-H) provides nearly optimal stratum boundaries (Cochran, 1977). The D-H method chooses boundary points such that equal intervals are created on the cum $\sqrt{f}$ scale (Dalenius and Hodges, 1959). Here, f is the density function for the stratification variable. Both Thomsen (1977) and Kish and Anderson (1978) use the cum $\sqrt{f}$ method separately on the density functions of Z_1 and Z_2 to create a bivariate D-H stratification.

Often an underlying linear model is assumed for y_i. The following model

$$y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i, \qquad (1)$$

is used by Anderson (1976) and Thomsen (1977) to develop properties of bivariate D-H stratification, where X_1 and X_2 are also the stratification variables, $E(e_i) = 0$, $Var(e_i) = \sigma_i^2 = c$ (a constant), and $Cov(e_i,e_j) = 0$ for $i \neq j$. A more general model,

$$y_i = \beta_o + \beta_1 X_{1i} + \beta_2 X_{2i} + e_i, \qquad (2)$$

is used by Roshwalb and Wright (1991), where X_1 and X_2 are two predictor variables of y_i, $E(e_i) = 0$, $Cov(e_i, e_j) = 0$ for $i \neq j$, and the variance is non-constant with

$$Var(e_i) = \sigma_i^2 = \sigma^2 \exp 2(\alpha_1 Z_{1i} + \alpha_2 Z_{2i}). \qquad (3)$$

The later study uses the model in (2) and in particular, the model for the residual variance in (3) to construct stratum boundaries. This method is an extension of model-based stratified sampling (MBSS) discussed in Wright (1983). The variables, Z_1 and Z_2, are used in the stratification through the model in (3), but they are not necessarily related to the predictor variables X_1 and X_2 in equation (2).

Statistical sampling, as used in inventory valuation, has two different audiences in the accounting-auditing environment. The first group are auditors from within the firm (internal auditors) whose task is to value the inventory. Their goal is to provide an efficient and economical estimate of the inventory worth for financial reporting purposes using methods acceptable to the second group, the external auditors. An external auditor is hired to evaluate whether the firm's stated assets and liabilities are accurate as stated. This may mean the external auditor actually samples the firms assets or just verifies whether procedures used by the internal auditor conform to the external auditor's standards. Although each group may consider the problem from a slightly different point-of-view, we assume that both groups share the same goal: to have a precise estimate of the firm's total value or total error.

One can study many different sampling methods in the context of this problem, since information is readily available for design and analysis. It is not unusual for a firm to track information on each item such as current dollar balances (book values), cumulative sales, forecasts of future demand, transaction activity, etc. With the exception of Roshwalb and Wright's (1991) study, all other auditing studies have restricted their examination to using book value as the sole stratification variable. Neter and Loebbecke (1975) examine the performance of the mean per unit, difference and regression estimators under simple random, Dalenius-Hodges stratified, and probability proportional to size sampling. They conclude that stratified difference or stratified regression estimation is an appropriate choice for this problem. Roshwalb, Wright and Godfrey (1987) consider only stratified difference estimation focusing their study on the efficiency D-H and MBSS on book value alone. Their results suggest that small to modest improvements in efficiency can be consistently found by using MBSS instead of D-H stratification. In Roshwalb and Wright (1991), volume is introduced as another stratification. Again, only the stratified difference estimator is studied using the following stratified design: MBSS based on book value alone, D-H based on book value alone, MBSS based on volume alone, D-H based on volume alone and the bivariate MBSS based on book value and volume designs. The results in this study suggest that significant improvements in efficiency can be found by using bivariate MBSS stratification.

In our study, we compare the efficiency of bivariate D-H stratification to the bivariate MBSS stratification. The univariate methods are included for comparison. As in Roshwalb, Wright and Godfrey (1987) and Roshwalb and Wright (1991), we consider only the stratified difference estimator. In Section 2, we discuss the stratification methods in greater detail. In Section 3, we provide a justification for the use of the difference estimator, and we provide an empirical comparison of the sample designs efficiencies for three actual inventory populations. Section 4 provides some discussion and conclusions.

2. STRATIFICATION METHODS

Without loss of generality, we consider only the goal of estimating the population total Y using a stratified design with H strata. In Section 1, we stated that previous studies of inventory valuation tend to suggest using the stratified difference estimator,

$$\hat{Y}_{st} = \Sigma_h N_h \bar{d}_h + \Sigma x_i \qquad (4)$$

with variance

$$\mathrm{Var}\,(\hat{Y}_{st}) = \Sigma_h N_h^2\,(1 - n_h/N_h)\,S_h^2/n_h. \qquad (5)$$

The notation for the inventory problem is: y_i is the actual value (audit value) for unit i; x_i is the book value (perpetually maintained) for unit i; $d_i = y_i - x_i$ is the difference for unit i; N_h is stratum size for stratum h; n_h is sample size for stratum h; $\bar{d}_h = (1/n_h)\,\Sigma\,d_i$ is the sample mean of the differences for stratum h; $D_h = (1/N_h)\,\Sigma\,d_i$ is the population mean of the differences for stratum h; and $S_h^2 = (N_h - 1)^{-1}\,\Sigma\,(d_i - D_h)^2$ is the sampling variance for stratum h.

For the difference estimator, the model in (2) has the book value for X_1 and assumes $\beta_1 = 1$ and $\beta_2 = 0$ for any other predictor variable. If β_1 is not 1 or β_2 is not 0, then the regression or ratio estimator is more appropriate. The form of the difference estimator in (4) and its variance in (5) hold for any stratified design.

The bivariate D-H stratified design uses the D-H rule on each stratification variable to create a cellular division of the population as depicted in Figure 1. The total number of strata is $H = H_1 \times H_2$. H_1 is the number of stratum boundaries determined using cum $\sqrt{f}$ rule on X_1 and H_2 is the number of stratum boundaries determined using the cum $\sqrt{f}$ rule on X_2. As pointed out in Kish and Anderson (1978), the number of strata can be very large for even small values of H_1 and small values of H_2. However, 4 or 5 strata boundaries on each dimension may be sufficient to provide substantial gains in efficiency, since most of the gains are seen in first several strata for univariate stratification.

Anderson (1976) proposes another method of stratification using the predicted values of y from an estimate of the model in (1). In this approach, the cell structure of the bivariate D-H approach is abandoned, and the regression model collapses the two variables into one. They find that this can be an effective method of stratification, but they are hesitant to endorse the method due to the data requirements and interpretability of the strata. Thomsen (1977) also uses the model in (1) to analyze the effects of the stratification variables in bivariate D-H stratification. In this study, bivariate D-H stratification improves the efficiency of sample design using the model in (1) as a basis for comparison. Both of these analyses assume that each stratification variable predicts the target variable, however, a stratifier may not always predict the levels of a target variable. As we will see in the next section, the volume stratifier does not predict the levels of the target variable. Volume does predict levels of variability in the target variable. This observation is consistent with Bethel's (1989) analysis that an optimal stratifier for a univariate design does not necessarily have to predict the level of the target variable as long as it predicts the variability of the target variable. For this reason, we do not further consider these methods.

In one way, the MBSS approach is similar to the regression method. The bivariate multiplicative model for the residual standard deviation, equation (2), collapses the two stratification variables into one stratification variable. In MBSS, the stratum boundaries are determined by creating a list of σ_i in ascending order, dividing the population into strata such that the within-strata sum of the σ_i are equal among strata. Under this approach, certain strata will have items with low expected variability and other strata will have items with high expected variability.

The efficiency of the sampling plan is determined by the sample allocation. Optimal (or Neyman) allocation apportions the sample with respect to the within-stratum standard deviations and yields the most efficient estimators (Cochran, 1977). This requires that each stratum S_h, or a good estimate for it, be known at the planning stage. Bethel (1989) reports that in an univariate environment, D-H stratification with optimal allocation approaches a lower variance bound defined by Godambe and Joshi (1965), i.e., asymptotically approaches an optimality criterion. Also, the MBSS design with optimal allocation in the same environment can be shown to attain the same lower bound. In the bivariate environment, it can be shown that univariate designs will not reach the lower variance bound, but the bivariate D-H and MBSS designs will.

In practice, the S_h are not known, and estimates for the S_h may not be available, therefore, optimal allocation is not possible. Instead of optimal allocation for bivariate D-H, Thomsen (1977) considers equal allocation, and Kish and Anderson (1978) considers proportional allocation. Wright (1983) suggests equal allocation for MBSS. For univariate designs, Bethel (1989) and Wright (1983) show that D-H stratification with proportional allocation does not approach the optimality criterion of the variance lower bound unless the residual variances are constant; they also show that MBSS stratification with equal allocation does approach the variance lower bound. D-H stratification with equal allocation only approaches the variance lower bound when unrealistic conditions are placed on the distribution of the stratification variable. Bethel's analysis can be extended to bivariate methods with similar results.

3. EMPIRICAL RESULTS

The inventory problem is a good area to study issues in finite population sampling methods. The population sizes tend to be moderate (on the order 1000 to 10,000 elements), and a great deal of information is maintained by the firms even for the complete population. For this study, three inventories are available with complete information on the book values, audit values and volume for each member of the population. A population member of an inventory is defined as an item category, for example, one item category may be the stock of one brand of 100-watt light bulbs and another category may be the stock of another brand 75-watt light bulbs. Each category is carried separately in the inventory records, and an end-of-period tally would report that the inventory has $20 of 75-watt light bulbs and $45 of 100-watt light bulbs left in stock. Each is considered a different item i with a value X_i. For the remainder of the paper, we'll refer to X_1 as the book value, X_2 as the volume, and Y as the audit value.

Figures 2, 3 and 4 suggest the difference model and the form of the residual variance. Figure 2 shows the relationship between the Audit and Book values. The relationship adheres to a 45 degree line with only slight and apparently random deviations. This implies book value is a strong predictor for the audit value and that the appropriate β is 1. Figure 3 shows the relationship between audit value and volume. Although some increasing relationship appears to exist between audit value and volume, it is very weak. Figure 4 shows the relationship between the difference between audit value and book value, d_i and the log of the stratification variable, either book value or volume. Figure 4 indicates that, once the relationship between audit and book is explained, volume does not predict the level of the audit. However, the fan shaped spread between the differences and the log of the book or volume in Figure 4 indicates that heteroscedasticity in the differences exists and it is a function of book value or volume. This analysis implies that we should try the log of the book value for Z_1 and the log of the volume for Z_2. As stated earlier, we'll assume the difference model with β_1 = 1 and β_2 = 0, i.e., equation (2) becomes

$$y_i = \beta_o + X_{1i} + e_i. \tag{6}$$

Estimates for the heteroscedastic linear model in (6) and (3) are reported in Roshwalb and Wright (1991). The parameter estimates for α_1 and α_2 have a maximum asymptotic standard errors of .06, which indicates that a bivariate model is reasonable. These results imply univariate sample designs, as well as bivariate D-H stratified designs with equal or proportional allocation, should not asymptotically achieve the variance lower bound. A bivariate MBSS stratified plan should asymptotically achieve the variance lower bound, and when the stratum variances are available, bivariate D-H stratified plans with optimal allocation and bivariate MBSS plans with optimal allocation should asymptotically achieve the variance lower bound.

Bethel's analysis uses the asymptotic variance lower bound as a basis for comparison, however, the lower bound is derived from the interrelationship between the sampling plan and the expected variance under the model (see Wright, 1983 for discussion). If the model's residual variance is misspecified, we are uncertain of the estimator's variance lower bound. We chose to examine each sampling plan under the classical sampling measure of efficiency. This analysis does not rely on the model specification for assessing the stratification's effectiveness. On the other hand, if the model is an accurate representation of our data, the sampling results should conform to those discussed in the previous section.

To examine the effect of the different design schemes, we determine the variance ratio for the H=20 stratified difference estimator using univariate D-H, univariate MBSS, bivariate D-H and bivariate MBSS designs. The efficiency of the stratified plan versus a simple random sampling plan is the variance ratio of the stratified estimator over the unstratified estimator, VR = $Var_{st}(\hat{Y}) / Var(\hat{Y})$. The definition of $Var_{st}(\hat{Y})$ remains the same as in (5). A small VR indicates that the stratification greatly reduced the estimator's variances, a VR greater than one indicates that the stratification increased the estimator's variance.

In the study, the stratum variances are available for planning, and optimum allocation is possible. Using optimum allocation, the efficiency of the sample design is solely a result of the stratifications effectiveness. Our paper promotes that by incorporating more information into the stratification, a greater reduction in the variance is possible. The VRs for the different methods using optimal allocation are presented in Table 1. A bivariate design should reduce the variance more than a univariate design. Only in one instance did a univariate design have a smaller VR than a bivariate design, the MBSS based volume alone for INV3 has a smaller VR than the VRs for the bivariate D-H designs. In this case, the model in the MBSS method provides enough additional information to yield a smaller VR than using two variables in a bivariate D-H plan. However, the VR for the bivariate MBSS design is lower than the MBSS design based on volume alone. In all cases, the bivariate MBSS designs VRs are less than those for the bivariate D-H designs. In some cases, the differences are small and in other case the differences are more substantial.

Optimum allocation is not feasible when the stratum variances are unknown or estimates are unavailable at the planning stage. Equal allocation is suggested by Wright (1983) for MBSS and Thomsen (1977) for bivariate D-H. Proportional allocation is suggested by Kish and Anderson (1978) for bivariate D-H. From the minimum variance argument, we expect that the VRs to increase for the bivariate D-H designs with equal allocation and proportional allocation. Wright (1983) indicates that the VRs for should increase slightly for the MBSS plans with equal allocation, however, these designs still asymptotically achieve the variance lower bound when the model holds.

Table 2 presents the results for MBSS and D-H designs with equal allocation. As expected each design has greater variance ratios than their counterpart with optimal allocation. The increase in the variance ratios are not substantial for some of the bivariate designs with equal allocation, see INV2 and INV3, which indicates the bivariate designs are effectively stratifying the population. Although not reported, the VRs for proportional allocation increased substantially for each inventory.

4. DISCUSSION

Bivariate MBSS and bivariate D-H stratification schemes are examined in this paper. The stratified difference estimator is chosen to observe the effects of stratification because: 1) The difference estimator is appropriate for the inventory valuation problem. 2) Unlike the ratio or regression estimators, the difference estimator is unbiased, and its variance has an exact and known form. 3) If a mean expansion, ratio, or regression estimator is appropriate for the problem, the effects of the stratification schemes are best observed by using that estimator. In the case of the ratio or regression estimator, the estimators are not unbiased and the variance is not exact, but the estimators would be asymptotically unbiased.

Stratification reduces the variance of an estimator by creating homogeneous subgroups of the population. Intuitively, when more relevant information is included in developing the strata, the subgroups should be more homogeneous. This may be accomplished by adding another stratification variable into the design or by including more structure through a model.

In our three inventory populations, the bivariate stratified designs were generally found to be more efficient than univariate designs. Bivariate MBSS stratified designs were always more efficient than any of the bivariate D-H stratified designs fòr that inventory. Minimum variance stratification theory suggests that each bivariate method discussed here is asymptotically as efficient as the others. Since the bivariate D-H method creates divisions first on each stratification variable and then combines these discrete divisions into the cellular bivariate design, the strata are not as refined as in the model-based approach which creates the divisions by using a continuum from the multiplicative model. Also, the bivariate D-H method should become more efficient as the total number of strata is increased. The results seem to suggest that the rate of convergence to the variance lower bound seems faster for the bivariate MBSS method than the bivariate D-H method.

Planning a MBSS design is not a simple process, the form of the model needs to be specified and reasonable values for the models parameters need to be ascertained. This may mean some statistical modelling using prior periods data or data from a pilot sample. For the extra effort, the MBSS method appears to provide more efficient estimators, better control of the number of strata, and less likelihood of empty or small strata. If one desires 20 strata, the MBSS method provides 20 strata. In the bivariate D-H method, the total number of strata is often less than the desired number of strata due to empty or small strata (Kish and Anderson, 1978). In an univariate D-H design, a small or empty stratum is collapsed into its neighbor stratum, but in the bivariate D-H approach, it is unclear how to handle this situation since there are more than one neighbor.

Another methodological issue exists. For the univariate D-H method, very large elements are often set aside and measured with certainty. For bivariate D-H, some units may be set aside on one dimension but not in the other dimension. There are no guidelines on how to handle this situation, we include these units in a certainty stratum. Under this approach, it is possible to have a large number of items fall in the certainty stratum. The MBSS method, whether univariate or bivariate, simply identifies which members should be set aside.

A criticism of some stratification methods is whether or not the strata have interpretable definitions. For the inventory problem, the bivariate D-H strata are defined by the size of the book values and the size of the volume. The bivariate MBSS method creates strata based on the size of the expected residual variance which is modelled by the size of the book value and the size of the volume. A high expected residual variance is a function of a high book value and high volume, a low expected residual variance is a function of a low book value or low volume, and an expected residual variance in the middle is a function of all other possible combinations.

To conclude, the results of this study indicate that bivariate stratification will produce better sample designs. The MBSS approach appears to better synthesize the information from two stratification variables and appears to have fewer methodological problems.

REFERENCES

Anderson, D. W. (1976) Gains from Multivariate Stratification, unpublished PhD. Dissertation, The University of Michigan.

Bethel, J. (1989), "Minimum variance estimation in stratified sampling," Journal of the American Statistical Association, 84, 260-265.

Cochran,W.(1977),Sampling Techniques, NY: Wiley.

Dalenius, T. and J.L. Hodges (1959), "Minimum Variance Stratification," Journal of the American Statistical Association, 54, 88-101.

Godambe, V.P. and V.M. Joshi(1965), "Admissibility and bayes estimation in sampling finite populations," The Annals of Mathematical Statistics, 36, 1707-1722.

Godfrey, J, A. Roshwalb and R.L. Wright (1984), "Model-Based Stratification in Inventory Cost Estimation," Journal of Business and Economic Statistics, 2, 1-9.

Kish, L. and D.W. Anderson (1978), "Multivariate and Multipurpose Stratification", Journal of the American Statistical Association, 73, 24-34.

Neter, J., and J.K. Loebbecke (1975), Behavior of Major Statistical Estimators in Sampling Accounting Populations: An Empirical Study New York: American Institute of Certified Public Accountants.

Roshwalb, A. and R. L. Wright (1991), "Using information in addition to book value in sample designs for inventory cost estimation," The Accounting Review, 66, 348-360.

Roshwalb, A., R.L. Wright and J. Godfrey (1987), "A New Approach for Stratified Sampling in Inventory Cost Estimation," Auditing: A Journal of Practice and Theory, 7, 54-70.

Thomsen, I. (1977) "On the effect of stratification when two stratifying variables are used," Journal of the American Statistical Association, 72, 149-153.

Wright, R.(1983)"Finite Population Sampling with Multivariate Auxiliary Information", Journal of the American Statistical Association, 78, 79-84.

Table 1: Design effects for univariate and bivariate MBSS and D-H stratified plans using optimal allocation (20 strata total)

	Model-based Stratified			Dalenius-Hodges					
	Book	Volume	Book	Book	Volume	Book and Volume			
Population	Alone	Alone	and Volume	Alone	Alone	10x2	5x4	4x5	2x10
INV1	0.304	0.394	0.282	0.369	0.418	0.290	0.287	0.295	0.345
INV2	0.271	0.285	0.249	0.336	0.309	0.288	0.284	0.285	0.284
INV3	0.223	0.166	0.126	0.252	0.174	0.184	0.181	0.181	0.183

Table 2: Design effects for univariate and bivariate MBSS and D-H stratified plans using equal (20 strata total)

	Model-based Stratified			Dalenius-Hodges					
	Book	Volume	Book	Book	Volume	Book and Volume			
Pop.	Alone	Alone	and Volume	Alone	Alone	10x2	5x4	4x5	2x10
INV1	0.332	0.500	0.317	0.410	0.479	0.456	0.538	0.615	0.516
INV2	0.537	0.375	0.338	0.652	0.423	0.321	0.315	0.315	0.313
INV3	0.413	0.195	0.159	0.424	0.199	0.198	0.198	0.199	0.206

Figure 1:

Cells for a H_1 by H_2 Bivariate D-H Stratified Design

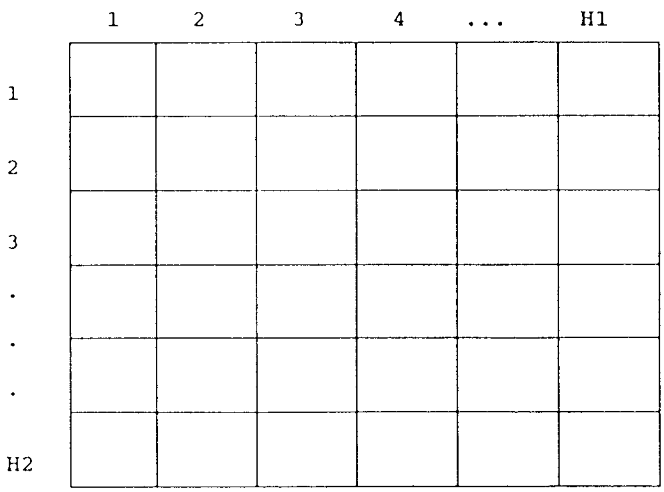

Figure 2:
Scatterplot of Audit versus Book

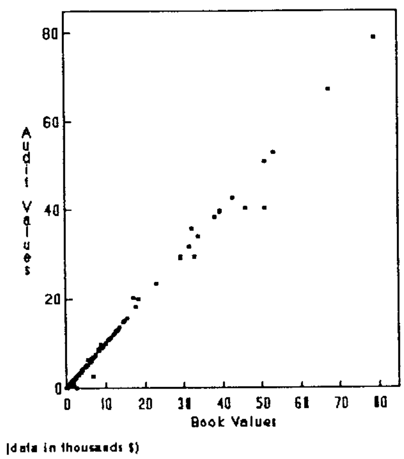

Figure 3:
Scatterplot of Audit versus Volume

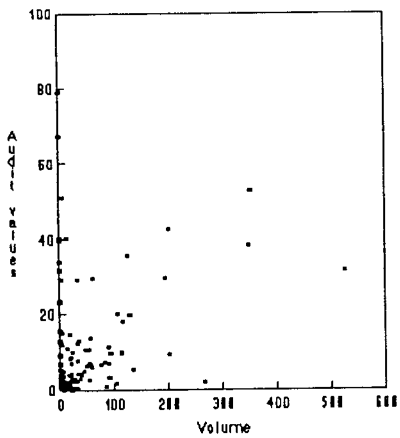

Figure 4:
Difference versus log of Stratification